基于语义位置和区域划分的兴趣点推荐模型

刘　辉[1,2]，万程峰[1]，吴晓浩[1]

(1. 重庆邮电大学 通信新技术应用研究中心，重庆 400065; 2. 重庆信科设计有限公司，重庆 401121)

摘　要：针对现有的位置社交网络研究工作对兴趣点相关的用户语义位置信息挖掘不够充分，且大多推荐算法忽略了兴趣点所在区域对推荐结果的影响，提出了一种新型兴趣点推荐模型(USTTGD)。首先采用分割时间的潜在狄利克雷分配主题模型(latent Dirichlet allocation，LDA)，基于签到记录中的语义位置信息挖掘时间主题下的用户时间偏好，然后将兴趣点所处区域划分为网格，以评估区域影响;接着应用边缘加权的个性化 PageRank (Edge-weighted Personalized PageRank，EwPPR)来建模兴趣点之间的连续过渡;最后将用户时间偏好、区域偏好和连续过渡偏好融合为一个统一的推荐框架。通过在真实数据集上实验验证，与其他传统推荐模型相比，USTTGD 模型在准确率和召回率上有了显著的提升。

关键词：位置社交网络；语义位置；兴趣点推荐；时间主题；区域影响

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2018.08.0541

Point-of-interest recommendation model based on semantic location and regional division

Liu Hui[1, 2], Wan Chengfeng[1], Wu Xiaohao[1]

(1. *Research Center of New Telecommunication Technology Applications, Chongqing University of Posts & Telecommunications, Chongqing* 400065, *China*; 2. *Chongqing Information Technology Designing Co. Ltd, Chongqing* 401121, *China*)

Abstract: According to the existing research work of location-based social network was not sufficient to mine the user semantic location information related to point-of-interest, Moreover, most recommendation algorithms ignored the influence of the region of point-of-interest on the result of recommendation. This paper proposed a new recommendation model of point-of-interest called USTTGD, first adopted the Latent Dirichlet Allocation(LDA) topic model of time division, based on the semantic location information in check-in records mined the user time preference under the time theme, then devided the region of point-of-interest into grids to evaluate the regional influence. Next, applied Edge-weighted Personalized PageRank(EwPPR) to modeling the successive transitions among point-of-interests. Finally, USTTGD fused user time preference, regional preference and successive transition preference into a unified recommendation framework. Experimental results on real-world datasets show that USTTGD achieves significantly enhance compared with other classical recommendation models on precision and recalling rates.

Key words: location-based social network; semantic position; point-of-interest recommendation; time theme; regional influence

0　引言

随着移动通信和 4G 网络技术的日渐成熟，基于位置的社交网络(location-based social network, LBSN)正变得前所未有的流行。社交网络用户以“签到”的形式在兴趣点(point-of-interest，POI)上分享他们的位置和体验。典型的 LBSN 网站有 Foursquare、Yelp、Twitter、Facebook、街旁、大众点评等，这些网站根据用户的历史签到数据来向用户推荐新的兴趣点(如公园、餐厅等)。兴趣点推荐服务不仅能给用户带来丰富的社交体验，同时还能为企业带来商业收益，提升商家知名度。因此，LBSN 兴趣点推荐逐渐成为推荐领域的研究热点。

位置推荐中现有的研究方法大多都是基于概率生成模型(PGM)。基于 PGM 的方法[1~3]从用户的签到记录中了解用户的潜在偏好，如潜在的空间偏好和局部偏好，利用马尔可夫链蒙特卡洛法和变分法近似推理，可以推导出潜在空间和局部变量的后验分布。基于矩阵分解的方法[4,5]可通过将用户-兴趣点矩阵分解为不同含义的潜在特征矩阵来预测用户的偏好。然而这些方法只能学习用户的静态偏好，不能捕获用户在一天中不同时段内的动态兴趣。基于概率生成模型的 USTTM 算法[7]通过考虑用户在不同的时段内所做的选择策略，从他们的历史签到数据中挖掘用户的动态时空主题。然而,它只能从用户的地理签到数据中的经纬度来捕获用户的时间偏好，没有任何语义解释。这些信息不足以捕获用户对不同类型位置的偏好。此外，兴趣点所处区域通常也会影响用户的选择，一般来说用户更倾向于选择前往著名景点区范围内附近的某个兴趣点进行签到。因此，位置推荐中兴趣点所处区域也是不可或缺的考虑因素。

综上所述，本文提出了一种统一兴趣点推荐模型，综合考虑了上述几种情境因素。本文的贡献主要为：a)提出了一

收稿日期：2018-08-06；**修回日期：**2018-09-21

作者简介：刘辉（1966-），男，四川仪陇人，正高级工程师，主要研究方向为通信网络新技术、电信系统业务；万程峰（1993-），男，湖北孝感人，硕士研究生，主要研究方向为个性化推荐、数据挖掘(1091734860@qq.com)；吴晓浩（1991-），男，湖北咸宁人，硕士研究生，主要研究方向为通信新技术应用、大数据、云计算.

种分割时间的 LDA 算法(STLDA)，将签到数据集划分为不同的时段并基于多个 LDA 模型进行训练，使用语义定位来发掘用户在不同位置上的时间偏好；b)在 STLDA 算法的基础上又提出了引入时间变量的主题挖掘算法(TVTM)，改善了 STLDA 中存在的数据稀疏性问题，可在单个训练模型中生成时间主题；c)通过将空间划分为网格来计算兴趣点的区域偏好；d)构建简化的兴趣点-兴趣点过渡图，并应用边缘加权的个性化 PageRank[7]计算图中每个兴趣点间的连续过渡偏好；e)在真实数据集上测试了该模型的准确率及召回率。实验结果表明，本文所提的兴趣点推荐模型各项性能指标均优于其他主流的推荐模型。

1　相关工作

随着位置社交网络的快速发展，兴趣点推荐开始迅速普及。本章主要回顾了与本文研究内容相关的最新研究工作的进展。兴趣点推荐中主流的推荐技术可分为三类:

a)基于概率生成模型的方法。利用概率生成方法来学习用户的偏好。Yin 等人[1]提出了学习用户区域和局部偏好的 JIM 模型。在他们的模型中，区域和局部偏好由潜在变量表示，这些潜在变量导致了用户在不同时间段访问不同地点的选择，潜在变量的后验分布由可见变量(如签到位置和时间)推断。这种模型的局限性在于用户可能会根据时间的影响作出调整。Hu 等人[3]利用稀疏编码技术对用户区域和时间偏好进行建模[4]，但未能考虑到时间影响。在他们的另一项研究中[2]，时间被建模为学习用户偏好的上游因素，但该模型面临的主要问题是：用户平均访问的兴趣点数过于稀疏。Wang 等人[5]讨论了一种新情况，即当用户前往一座新城市时，他们往往会动态地改变他们的兴趣和行为，而用户访问的大部分地点距离他们的家乡很远。为了应对这一挑战，他们提出了“地理鼠标器”来建模两种不同类型的偏好：用户的本地偏好和旅行偏好。Liu 等人[7]提出 USTTM 模型来挖掘时空主题，可以在不同的时间段内找到用户对应的偏好。但该模型仅仅考虑了地理位置这一种情景因素，而没有借助语义位置信息。Liu 等人[10]构建了一种基于低秩图的方法来学习用户的连续时间模式，可以根据距离用户上次访问的时间间隔来向用户推荐位置，但该方法只能向用户推荐已访问过的位置。

b)基于矩阵分解的连续兴趣点推荐。Li 等人[4]利用矩阵分解的方法来寻找潜在的用户和位置特征矩阵。Wu 等人[5]采用了一种概率矩阵分解方法，综合考虑了用户选择某个地方的地理和时间因素。通过将不同的因素整合到一个概率生成模型中来建模用户对某个位置的兴趣。文献[6]提出了一种考虑用户偏好、近邻性和社交关系的协同过滤推荐方法。根据用户的相似度来进行位置推荐。但由于用户-兴趣点矩阵数据过于稀疏，导致推荐结果不够准确。Mao 等人[9]提出了一种结合用户偏好和空间影响的混合推荐系统。Lian 等人[11]也考虑了邻近区域的空间影响，然后应用矩阵分解方法来推荐兴趣点。Chen 等人[12]首先提出了连续兴趣点推荐的概念。将合并因式分解个性化马尔可夫链(FPMC)[12]融入 LBSN 中。Feng 等人[13]提出了一种度量嵌入(ME)模型，以观察在连续签到中潜在空间的关系。同时还考虑了用户和兴趣点之间的交互，以获得更好的推荐性能。

c)基于核密度估计的方法。利用核密度估计[13]来查找用户访问不同位置的时间间隔模式。Lichman[14]等人采用了混合核密度评估法在个体用户层面上寻找空间密度。文献[15]采用带有固定核宽的核密度估计方法来为基于每个用 户的经纬度坐标兴趣点地理位置的签到分布建模。Lian 等人[16]从二维核密度估计的角度来刻画空间聚集效应，并将其整合进矩阵分解模型中，通过对空间聚集效应的建模可以有效缓解用户—兴趣点矩阵的数据稀疏性问题。

总体来说，现有的研究工作对用户的语义位置信息考虑不够充分，同时忽略了连续兴趣点推荐中的区域影响因素，导致算法推荐性能往往较低。因此，本文提出了一种 USTTGD 模型，该模型的优势在于能通过分析用户签到记录中的历史语义位置信息，基于挖掘出的时间主题推断用户当前和未来最可能前往的兴趣点的位置。同时考虑到了区域影响因素和兴趣点间连续过渡因素，提升了推荐结果的准确性。实际上，现实生活中用户的行为偏好本身就受到多方面情景因素的影响，因此，本文所提模型更能反映真实的场景，贴合用户的实际行为。

2　基于语义位置信息的建模

2.1　问题描述

位置社交网络中的兴趣点推荐是通过分析用户的历史签到记录数据，向用户推荐之前未访问过的、可能感兴趣的位置。LBS 中应包含用户集 $U=\{u_1,u_2,\ldots,u_n\}$、兴趣点集 $L=\{l_1,l_2,\ldots,l_m\}$ 以及用户在兴趣点上的历史签到记录集。LSBN 体系结构 $G=\{U,L,D\}$ 如图 1 所示。包含了若干用户、兴趣点以及三类关系—用户间的社交关系，兴趣点间的关联关系及用户与兴趣点间的签到关系。实际上通过分析用户历史签到记录就能提取这三种关系。本文所提的推荐模型将通过融合多种不同源的数据，预测用户对未访问过的兴趣点的偏好，进而为其推荐兴趣点。

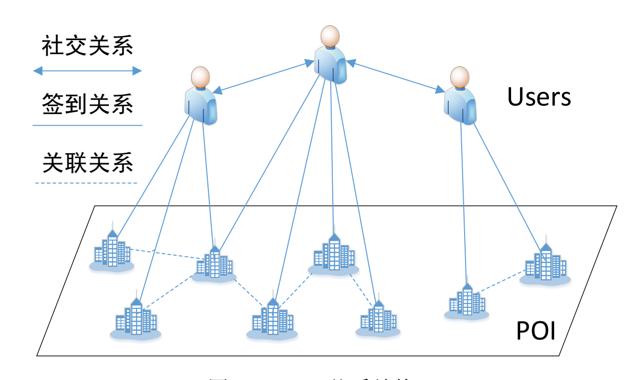


图 1　LBSN 体系结构

Fig. 1.　LBSN architecture

表 1 列出了本文算法所涉及到的符号及相关含义。

表 1　符号描述

Table 1　Symbols description

符号	含义
U,L,D	用户集合;兴趣点集合;签到记录集合
u,l,t	用户 $u\in U$;兴趣点 $l\in L$;签到时间 t
g_l	兴趣点 l 的地理位置,表示为经纬度坐标
W_l	关于兴趣点 l 的评论集合
W_u	用户 u 对其访问过的所有兴趣点的评论集合
Υ	兴趣点 l 在用户间的流行度
D_u	D 中属于用户 u 的所有签到记录集合

本文算法的推荐架构图如图 2 所示。

2.2　STLDA 算法

潜在的狄利克雷分配(LDA)[8]是一种著名的概率统计生成模型，可以在大型语料库中找到不同的主题集合。如图 3 所示，STLDA 算法的每个时间段都是 LDA 模型。LDA 的输入是一个文本集 D 和主题数 K，输出将是每个文本的主题分

布向量 θ,以及 K 个主题的主题词语分布 ϕ。w 代表文档中的单词。X 是一个潜在的变量，它表示在语料库 D 中每个文档里的每个单词 w 的主题索引，每个文档中词语的数量由 N_D 表示。每个文档被视为是一种多项分布的主题,每个主题被建模为词汇表的多项分布,θ 是每个文档的多项分布，ϕ 是每个主题的多项分布。α 和 β 为分别对 θ 和 ϕ 使用狄利克雷先验的超参数。为了确定文档中每个单词的潜在主题索引，采用 Gibbs 抽样方法，利用式（1）推断潜在变量 X，从而了解用户的偏好。

$$p(x_{d,i}=x\mid x_{d,\neg i},W_{d,\neg i},\alpha,\beta)\propto \frac{n^{x}_{d,\neg i}+\alpha_k}{\sum_{k=1}^{K}(n^{k}_{d,\neg i}+\alpha_k)}\times\frac{n^{h}_{x,\neg i}+\beta_h}{\sum_{h=1}^{H}(n^{h}_{x,\neg i}+\beta_h)} \tag{1}$$

其中：$x_{d,i}$ 是文档 d 中单词 i 的主题索引，$n^{x}_{d,\neg i}$ 是文档 d 中除了第 i 个单词外被指定为主题索引 x 的其他单词数。$n^{h}_{x,\neg i}$ 是被指定为主题索引 x 的索引为 h 的单词数。在对文档集中的每个单词分配主题索引并进行多次迭代之后，可以计算出每个文档的主题分布 θ 和每个主题的主题词语分布 ϕ，计算公式如下:

$$\hat{\theta}_{d,k}=\frac{n^{k}_{d}+\alpha_k}{\sum_{k=1}^{K}(n^{k}_{d}+\alpha_k)} \tag{2}$$

$$\hat{\phi}_{k,h}=\frac{n^{h}_{k}+\beta_h}{\sum_{h=1}^{H}(n^{h}_{k}+\beta_h)} \tag{3}$$

其中：n^{k}_{d} 是文档 d 中主题 k 下的单词数, n^{h}_{k} 是文档集中中主题 k 下索引为 h 的单词数，α_k 和 β_h 分别是 θ 和 ϕ 的狄利克雷先验分布。

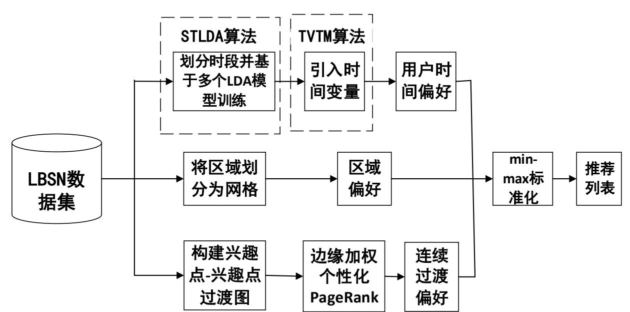


图 2　基于语义位置和区域划分的推荐架构图

Fig. 2　Recommendation architecture diagram based on senmatic location and regional devision

为了了解用户对兴趣点的时间偏好，将签到数据集按照签到时间划分为不同的时段，然后利用划分好的签到数据对 LDA 模型进行训练。这样就可以得到用户的时间偏好。对签到数据集每一个划分的时段，将每个用户视为文档，用户访问的语义位置是文档的内容。通过对时间的划分,可以了解每个用户在每个时段 t 对所呈现的主题向量 θ_d 的偏好。

2.3　TVTM 算法

STLDA 通过将签到数据集划分为不同的时间段来了解用户的时间偏好，但这种方法可能导致对用户的推荐结果不够准确，因为每个时间段中的训练数据集将会变得更稀疏。为了改善这个问题，提出了通过考虑时间变量来修改 LDA 的 TVTM 算法。如图 4 所示，TVTM 模型通过引入时间变量 t 来建模用户的时间偏好。具体实现方法是为每个用户生成 T 个 θ 的狄利克雷先验分布。离散时间变量为 s，s 是一个选择不同的主题分布变量 θ 的时间指标,θ 和 ϕ 的先验分布定义式如下:

$$p(\theta_{u,t}\mid\alpha)\propto\prod_{k=1}^{K}\theta_{u,t,z}^{\alpha_k-1} \tag{4}$$

$$p(\varphi_k\mid\beta)\propto\prod_{h=1}^{H}\varphi_{k,h}^{\beta_h-1} \tag{5}$$

其中：α 和 β 分别是 $\theta_{u,t}$ 和 ϕ_k 的狄利克雷先验分布的超参数。对用户数据集 U 中的每个用户，N_u 代表用户 u 的签到记录数，W_l 是由用户 u 访问的位置 l 所包含的评论词语数。潜在变量 X 是在 W_l 中每个语义词的主题索引,ϕ 是该词的主题分布。TVTM 的生成过程如下:a)从狄利克雷先验分布 β 中获取 K 项变量 ϕ;b)从 U 中每个用户上获取 T 项变量 θ;c)对每个用户 $u\in U$,应当满足签到记录 $d^{th}\in\{1,\ldots,N_u\}$；在 d^{th} 中将签到时间 t 离散化为 s；从 W_l 中每个单词上获取主题 $X\sim Multinomail(\theta_s)$。

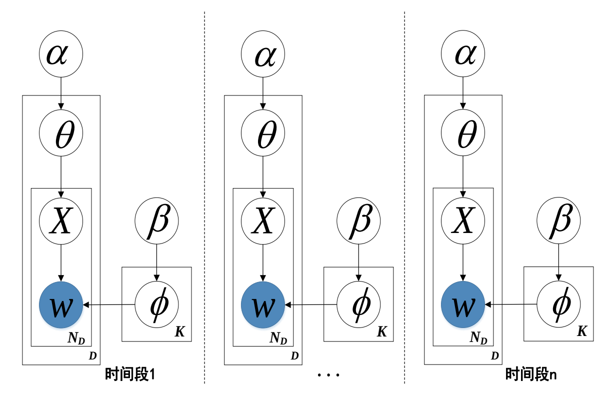


图 3　STLDA 算法模型

Fig. 3　The algorithm model of STLDA

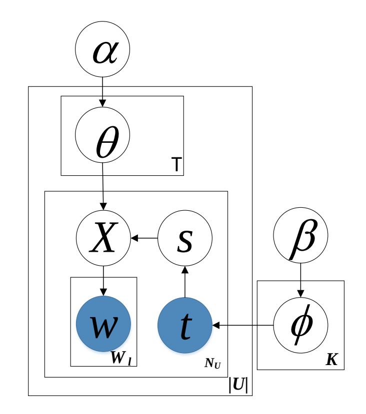


图 4　TVTM 算法模型

Fig. 4　The algorithm model of TVTM

1）用户时间偏好的学习

由于位置 l 有评论词汇集 W_l，可以从签到记录集 D_u 中提取一个评论集 W_u。将每个用户视为一个文档，通过从 TVTM 模型中学习到的主题分布来表示用户的偏好。由于当人们处于一天中的不同时间间隔时，往往会有不同的活动。因此人们决定去某地时，时间应该被视为一个重要因素。TVTM 中将时间 t 建模为用户在不同时间点的偏好指标。签到时间 t 将是一个连续时间变量并且会导致无限多的用户主题分布 θ_u。为了解决这个问题，将时间离散化为 T 部分，变量 s 代表 t 的时间索引。根据 LDA 模型,从主题 x 中选出的单词 w 有关于 ϕ_x 的多项式分布。从 θ_s 中选出的主题 x 表示用户在时间间隔 s 下的主题偏好。

令 $\Theta=\{\theta_u\}_{u=1}^{|U|}$ 为每个用户的主题分布张量,其中 θ_u 是一个

关于时间间隔和主题索引的矩阵。$\Phi=\{\varphi_k\}_{k=1}^{K}$ 表示在不同的主题 k 下的词汇分布，$\omega=\{W_l\}_{l=1}^{L}$ 代表整个评论集，变量的联合分布表示公式如下：

$$p(x,t,s,\omega,\Theta,\Phi\mid\alpha,\beta)=p(x\mid\Theta)p(\Theta\mid\alpha)p(\omega\mid z,\Phi)p(\Phi\mid\beta)p(s\mid t) \tag{6}$$

2）TVTM 算法的参数推导

本节的目的是得到 x,t,s,ω 所提供的 Θ 和 Φ 的后验分布，也就是说需要计算 $P(\Theta,\Phi|\ x,t,s,\omega)$。由于难以计算边缘分布 $P(\ x,t,s,\omega\ |\alpha,\beta)$，故精确推出 $P(\Theta,\Phi|\ x,t,s,\omega)$是不可能的。因此，考虑应用马尔可夫链蒙特卡罗(MCMC)方法折叠 Gibbs 抽样[15]进行近似推导。选择一个 θ 和 ϕ 的狄利克雷共轭先验时，可以集成 Θ 和 Φ 来简化抽样过程。

$$p(x_{u,i}=x\mid x_{u,\neg i},\omega_{u,\neg i},t,s,\alpha,\beta)\propto \frac{n_{u,s,\neg i}^{x}+\alpha_x}{\sum_{k=1}^{K}(n_{u,s,\neg i}^{k}+\alpha_k)}\times\frac{n_{x,\neg i}^{h'}+\beta_h}{\sum_{h=1}^{H}(n_{x,\neg i}^{h}+\beta_h)} \tag{7}$$

根据折叠 Gibbs 采样的程序，需要迭代更新完整的条件分布，如式(7)所示,迭代获取 W_u 中出现的每个单词的新主题。定义 $n_{u,s,\neg i}^{x}$ 为第 u 个用户观察到的主题 x 的次数，不包括 W_u 中第 s 个时间索引下的第 i 个单词。同样, $n_{k,s,\neg i}^{h}$ 是第 s 个时间间隔中分配给主题 k 下的单词 h 的数量。经过足够多次的迭代后，马尔可夫链逐渐收敛到一个平稳分布，每个用户的主题分布和单词分布可由下面的式（8）和（9）计算得到，设置 α= 50 / K，β= 0.01。

$$\hat{\theta}_{s,u,k}=\frac{n_{u,s}^{k}+\alpha_k}{\sum_{k=1}^{K}(n_{u,s}^{k}+\alpha_k)} \tag{8}$$

$$\hat{\varphi}_{k,h}=\frac{n_{k}^{h}+\beta_h}{\sum_{h=1}^{H}(n_{k}^{h}+\beta_h)} \tag{9}$$

2.4　用户时间偏好

LBSN 中先前的研究工作大都以静态的方式向用户推荐位置，而基于时间主题的建模则是在向用户推荐兴趣点时考虑到用户的时间偏好。具体来说，根据给定的用户签到记录组合 (u,l,t)，对于 L 中的每个兴趣点 l，计算出用户 u 访问 l 的可能性，同时兴趣点评分也被视作是用户选择访问位置的一个重要因素。u 对 l 的时间偏好定义为 u 在时间 t 访问兴趣点 l 的概率，其计算公式如下。

$$p_{u,l,t}^{user}=p(l\mid W_l,\gamma_l,u,t,\hat{\theta},\hat{\varphi})=\frac{p(l,W_l,\gamma_l\mid u,t,\hat{\theta},\hat{\varphi})}{\sum_{l'}(p(l',W_{l'},\gamma_{l'}\mid u,t,\hat{\theta},\hat{\varphi}))} \tag{10}$$

其中：$p(l,W_l,\gamma_l\mid u,t,\hat{\theta},\hat{\varphi})$ 的计算如式（11）所示，γ_l 是[0,1]规范化。

$$p(l,W_l,\gamma_l\mid u,t,\hat{\theta},\hat{\varphi})=Beta(\gamma_l,1)\times\sum_{k=1}^{K}P(x\mid u,\hat{\theta})\prod_{w\in W_l}p(w\mid x,\hat{\varphi})^{\frac{1}{|W_l|}} \tag{11}$$

3　基于区域划分的建模

3.1　区域偏好

考虑到兴趣点所处区域也会对推荐结果产生影响。引入网格思想。如图 5 所示，将空间划分为几个网格单元。设 l 为用户 u 的当前位置。以 l_c 为圆心，r 为半径的圆圈覆盖内的网格单元，称为 u 的近邻网格单元，令 $num(D_{g_i})$ 是一个网格单元格 g_i 中所有兴趣点的签到记录数量的总和。使用下列公式来测量网格单元格 g_i 的流行度。

$$g_i^{\gamma}=\frac{num(D_{g_i})}{\sum_{\forall g\in G_b}num(D_g)} \tag{12}$$

其中 G_b 是 u 的近邻网格单元集合。

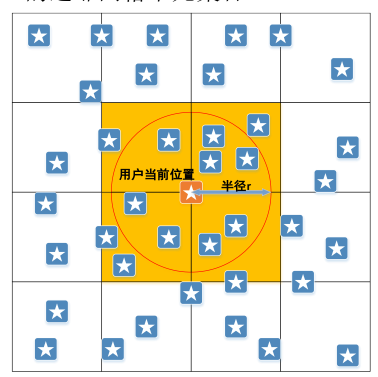


图 5　网格区域划分

Fig. 5　Grid regional devision

当用户 u 在网格单元 g_i 中的一些兴趣点上有大量签到记录时,表明网格单元 g_i 很可能是用户 u 喜欢的区域,同时 u 在网格单元 g_i 中的其他兴趣点上签到的可能性也会很高。令 $num(D_{g_i}^{u})$ 为用户 u 在网格单元 g_i 中所有兴趣点上签到记录的数量。然后利用下列公式来计算这种偏好。

$$g_i^{u}=\frac{num(D_{g_i}^{u})}{\sum_{\forall g\in G_b}num(D_g^{u})} \tag{13}$$

由于用户更倾向于选择在靠近用户当前位置的兴趣点上进行签到。所以认为用户倾向于在用户当前所处的网格单元内的兴趣点上签到,故定义一个偏好变量 g_i^{c},当用户当前位置在 g_i 内时，变量值为 1，否则为 0。

然后将上述三种偏好变量组合在一起。由下列线性函数表示网格单元格 g_i 的偏好得分。

$$Gps(g_i)=\lambda g_i^{\gamma}+\delta g_i^{u}+\varepsilon g_i^{c},\quad 0\le\lambda,\delta,\varepsilon\le1;\lambda+\delta+\varepsilon=1 \tag{14}$$

最终，u 对兴趣点 l 的区域偏好计算公式如下:

$$p_{u,l,t}^{reg}=\frac{Gps(g_i)}{\sum_{\forall g\in G_b}Gps(g)} \tag{15}$$

3.2　连续过渡偏好

在本节中，利用兴趣点—兴趣点过渡图来建模连续签到的关系。令(u,l,t)表示用户 u 在时间 t 时在兴趣点 l 上的签到记录。兴趣点-兴趣点过渡图定义如下。

定义 1　用户 u 的签到记录为$(u,l_1,t_1),(u,l_2,t_2),\ldots,(u,l_n,t_n)$，其中 $t_1\le t_2\le\ldots\le t_n$。如果 $t_{i+1}-t_i<\tau$，τ 表示连续签到的时间间隔，则认为关于用户 u 从兴趣点 l_i 到 l_{i+1} 之间有一个连续的过渡。

定义 2　用户 u 的兴趣点-兴趣点过渡图是一个有向图 $G=(L,E)$,其中 L 是所有兴趣点的集合，E 是 L 中所有的连续过渡集合。如果在 u 的历史签到记录中存在一个从 l_i 到 l_j 的连续过渡，那么 E 中将存在一条定向边(l_i,l_j)。(l_i,l_j)边的权值定义如下。

$$Ew(l_i,l_j)=\frac{Tran(l_i,l_j)}{\sum_{\forall l\in L}Tran(l_i,l)} \tag{16}$$

其中：$Tran(l_i,l_j)$为在 u 的历史签到记录中从 l_i 到 l_j 的连续过渡数。

由于只有与用户 u 当前位置距离小于 d 的兴趣点作为候

选兴趣点被推荐，故再提取一个 G 的子图，命名为兴趣点-兴趣点的过渡图 $G'=(L',E')$，通过移除不在 u 所处的网格单元或近邻网格单元格中的兴趣点。然后使用边缘加权个性化 PageRank（EdgePPR）[15]来计算 G' 中每个兴趣点的 EdgePPR 值。u 对 l 的连续过渡偏好定义为 l 的归一化 EdgePPR 评分，可以通过以下公式获得。

$$p_{u,l,t}^{suc}=\frac{EdgePPR(G',l)}{\sum_{\forall l'\in L'}EdgePPR(G',l')} \tag{17}$$

3.3 规范化统一模型

本文将利用 min-max 标准化来处理原始数据，将用户时间偏好、区域偏好和连续过渡偏好规范化如下：

$$\begin{aligned} S_{u,l,t}^{user}&=\frac{p_{u,l,t}^{user}-\min^{user}}{\max_u^{user}-\min_u^{user}}\\ S_{u,l,t}^{reg}&=\frac{p_{u,l,t}^{reg}-\min^{reg}}{\max_u^{reg}-\min_u^{reg}}\\ S_{u,l,t}^{suc}&=\frac{p_{u,l,t}^{suc}-\min^{suc}}{\max_u^{suc}-\min_u^{suc}} \end{aligned} \tag{18}$$

其中：$\max_u^{user}/\min_u^{user},\max_u^{reg}/\min_u^{reg},\max_u^{suc}/\min_u^{suc}$ 分别为 L' 中所有兴趣点上最大/最小的用户时间偏好、区域偏好以及连续过渡偏好。

综上所述，提出一种线性统一生成框架来集成这几种情景信息，最终用户 u 对兴趣点 l 的的总体偏好评分由下列公式可得。

$$\begin{aligned} &S_{u,l,t}=\sigma S_{u,l,t}^{user}+\rho S_{u,l,t}^{reg}+\eta S_{u,l,t}^{suc}\\ &0\le\sigma,\rho,\eta\le 1;\sigma+\rho+\eta=1 \end{aligned} \tag{19}$$

3.4 参数推导

TVTM 算法中详细参数推导见算法 1。

算法 1　TVTM 中的 Gibbs 抽样

输入：用户签到数据集 D，主题数 K，时间间隔数 S，

迭代次数 I，预处理时间 I_b，样本滞后时间 I_g，先验分布 α，β

输出：目标参数 $\hat{\theta},\hat{\varphi}$

创建计数变量 $n_k^h,n_{u,s}^k,\theta^{sum},\varphi^{sum}$，全初始化为 0

for　$u\in U$　do

　for　$w\in W_u$　do

　　为 w 随机分配话题，更新计数变量 $n_k^h,n_{u,s}^k$

　end

end

创建变量　count=0;

for　迭代次数从 1 到 I　do

　for　u∈U　do

　　for　w∈W_u　do

　　　利用式（7）更新主题分配，更新 $n_k^h,n_{u,s}^k$

　　end

　end

　if(iteration>I_b && iteration%I_g==0)　then

　　count=count+1;

　　$\theta_{s,u,k}^{sum}=\frac{n_{u,s}^k+\alpha_k}{\sum_{k=1}^K(n_{u,s}^k+\alpha_k)}$

　　$\varphi_{k,h}^{sum}=\frac{n_k^h+\beta_h}{\sum_{h=1}^H(n_k^h+\beta_h)}$

　end

end

返回参数 $\hat{\theta}=\frac{\theta^{sum}}{count},\hat{\varphi}=\frac{\varphi^{sum}}{count}$

求解线性无约束最优化问题的常用办法是梯度下降法，因此，本文将对式（19）采用多元线性回归进行处理，转换表达式如下：

$$f(x)=f_c(x)=c_1x_1+c_2x_2+c_3x_3 \tag{20}$$

其中满足：

$$\begin{aligned} &f(x)=S_{u,l,t}\\ &c_1x_1=\sigma S_{u,l,t}^{user},c_2x_2=\rho S_{u,l,t}^{reg},c_3x_3=\eta S_{u,l,t}^{suc} \end{aligned} \tag{21}$$

对于损失函数，有

$$L(c)=L(c_1,c_2,\dots c_n)=\frac{1}{2}\sum_{i=1}^n(f_c(x^{(i)}-y^{(i)}))^2 \tag{22}$$

接着便可利用梯度下降法得到最优化参数：

$$\sigma=c_1-\frac{1}{n}\sum_{i=1}^n(f_c(x^{(i)}-y^{(i)}))x_1^{(i)} \tag{23}$$

$$\rho=c_2-\frac{1}{n}\sum_{i=1}^n(f_c(x^{(i)}-y^{(i)}))x_2^{(i)} \tag{24}$$

$$\eta=c_3-\frac{1}{n}\sum_{i=1}^n(f_c(x^{(i)}-y^{(i)}))x_3^{(i)} \tag{25}$$

4　实验结果及分析

4.1　实验数据集

本文实验使用了两个真实签到数据集，Foursquare 和 Gowalla。为了保证实验的有效性，去除签到记录数少于 5 的用户以及被签到数少于 80 的兴趣点，最终得到的 Foursquare 数据集包含 3 067 个用户的 180544 条签到记录，其中兴趣点数量为 27 564 个。Gowalla 数据集包含 6304 个用户的 808 172 条签到数据，兴趣点数量为 53 827 个。对 Foursquare 和 Gowalla 这两个数据集中的每位用户随机选取其中 75% 的签到数据作为训练集，余下 25%的签到数据作为测试集。

4.2　评价指标

本文采用两个在推荐算法中应用较为广泛的评价指标：准确率 precision@N 以及召回率 recall@N。N 代表最终推荐结果 Top-N 下的推荐数量，准确率表示算法推荐结果与用户反馈的契合程度，能够反映推荐的准确性。召回率则被用来评估算法的执行效率，体现的是用户偏好的推荐对象能被推荐的概率，反映推荐的全面性。计算方法为

$$precision@N=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|R(u)|} \tag{26}$$

$$recall@N=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|} \tag{27}$$

其中：R(u)表示推荐算法在执行训练集后得到的兴趣点推荐列表，T(u)则表示用户在测试集上的实际签到过的兴趣点列表。

4.3　实验参数选取

本节旨在选取能使 USTTGD 模型性能最优化的参数。本文设置主题数 K=50，时间段为 4，α= 1，β= 0.01，根据 3.4 节所叙述的参数推导方法，选择此时的最优化参数作为本文实验的参数。这些参数的值如表 2 所示。

表 2　实验参数值

Table 2.　Experimental parameter value

数据集	λ	δ	ε	σ	ρ	η
Foursquare	0.2	0.3	0.5	0.4	0.2	0.5
Gowalla	0.3	0.2	0.6	0.4	0.3	0.2

4.4　实验性能比较

为了验证本文所提出的推荐模型 USTTGD 的性能，将它与下列几种推荐算法进行实验对比。所对比的推荐算法详细特征如表 3 所示。

表 3　比较的推荐算法

Table 3　Recommendation algorithm for comparison

算法(简称)	算法描述
ULR[8]	基于用户的签到记录数据，融入了地理信息来进行推荐。
UGPLR[13]	基于用户的签到记录数据，结合了地理信息和兴趣点间的连续过渡因素来进行推荐。
FPMCLR[14]	基于用户的签到记录数据，结合地理信息和个性化马尔科夫链的因式分解来进行推荐。
USTTM[7]	基于概率生成模型，仅利用用户签到记录中的地理信息和时间信息来捕获用户时间偏好，无任何语义解释。
USTTGD	本文所提的推荐算法，基于概率生成模型，综合考虑了用户历史签到行为、时间信息、语义位置信息、地理信息、连续过渡影响。

实验 1　不同兴趣点推荐数量下的算法结果对比

本节实验主要观察各种算法在不同的兴趣点推荐数量(TOP-N)下的结果。将半径 r 和网格边长分别设为 1km,0.5km，如图 5~8 所示，图中的横轴 N 代表了所推荐的兴趣点个数，纵轴 *precision@N* 和 *recall@N* 分别代表在不同推荐兴趣点数量时各推荐算法对应的准确率及召回率。实验中分别设置 N=5、10、15、20，算法中的其余参数均设为满足算法性能最优化时的参数值。本节给出了各类算法在不同兴趣点推荐数量下准确率及召回率的比较结果。

表 4　Foursquare 数据集中不同 N 值下的推荐性能对比表

Table 4　Recommendation performance comparison table for different N values in the Foursquare dataset

算法	precision@N				recall@N			
	@5	@10	@15	@20	@5	@10	@15	@20
ULR	0.038	0.033	0.03	0.027	0.115	0.14	0.198	0.226
UGPLR	0.046	0.04	0.034	0.031	0.155	0.215	0.253	0.285
FPMCLR	0.058	0.053	0.046	0.041	0.163	0.223	0.256	0.288
USTTM	0.06	0.055	0.047	0.042	0.165	0.225	0.259	0.292
USTTGD	0.063	0.058	0.049	0.045	0.17	0.228	0.263	0.305

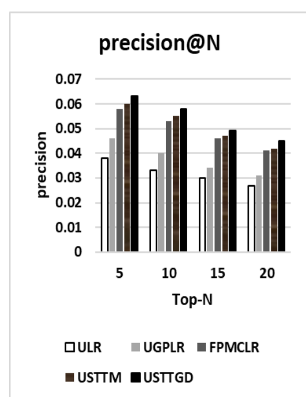


图 6　基于 Foursquare 的准确率实验对比图

Fig. 6　Precision experimental comparison based on Foursquare

图 7　基于 Foursquare 的召回率实验对比图

Fig. 7　Recalling rate experimental comparison based on Foursquare

从实验 1 结果可以看出：

a)随着 top-N 值的增加，各类算法的准确率均会有所下降，这是因为随着推荐数量的增多会增加模型的时间复杂度。但召回率均有所提升。

b)UGPLR、FPMCLR、USTTM、USTTGD 与 ULR 相比在性能上均有了较为明显的提升。这说明考虑多种情境因素相比仅考虑单一地理因素，对传统推荐算法性能的提升会起到更明显的作用。

c)USTTGD 相比 USTTM，在算法性能上有了进一步的提升，这说明基于用户语义位置信息的建模能获得更精确的推荐效果，而 USTTM 仅仅只是利用简单的地理坐标来挖掘时间主题，但相对于另外三种推荐算法，在推荐性能上也具有了足够的优势。

表 5　Gowalla 数据集中不同 N 值下的推荐性能对比表

Table 5　Table of recommendation performance comparisons for different N values in Gowalla dataset

算法	precision@N				recall@N			
	@5	@10	@15	@20	@5	@10	@15	@20
ULR	0.047	0.032	0.022	0.016	0.21	0.24	0.27	0.31
UGPLR	0.065	0.053	0.047	0.041	0.27	0.31	0.33	0.37
FPMCLR	0.068	0.056	0.051	0.045	0.29	0.33	0.35	0.4
USTTM	0.069	0.057	0.051	0.046	0.3	0.35	0.38	0.42
USTTGD	0.073	0.061	0.054	0.049	0.33	0.38	0.42	0.46

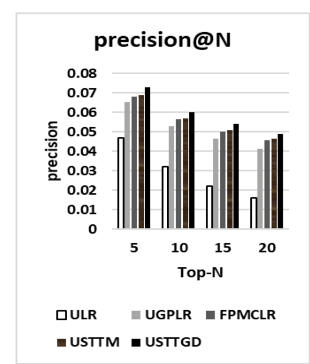


图 8　基于 Gowalla 的准确率实验对比图

Fig. 8　Precision experimental comparison based on Gowalla

图 9　基于 Gowalla 的召回率实验对比图

Fig. 9　Recalling rate experimental comparison based on Gowalla

实验结果表明，本文算法 USTTGD 相对于另外四种推荐算法，无论是在准确率还是召回率上，算法性能明显更好。

实验 2　不同签到时间间隔下的算法结果对比

本节实验主要观察各种算法在不同的连续签到时间间隔(τ)下的结果。将半径 r 和网格边长分别设为 1km,0.5km，将 τ 值分别设为 1、2、3、6 小时，实验结果如表 6、7 及图 10~13 所示。

表 6　Foursquare 数据集中不同 τ 值下的推荐性能对比表

Table 6　Table of recommendation performance comparisons for different τ values in Foursquare dataset

算法	评价指标	τ(小时)			
		1	2	3	6
ULR	precision@10	0.028	0.025	0.023	0.019
	recall@10	0.22	0.18	0.17	0.13
UGPLR	precision@10	0.041	0.035	0.03	0.024
	recall@10	0.26	0.23	0.2	0.18
FPMCLR	precision@10	0.046	0.039	0.033	0.027
	recall@10	0.27	0.24	0.21	0.19
USTTM	precision@10	0.049	0.042	0.035	0.028
	recall@10	0.29	0.26	0.23	0.2
USTTGD	precision@10	0.054	0.045	0.039	0.033
	recall@10	0.34	0.29	0.27	0.25

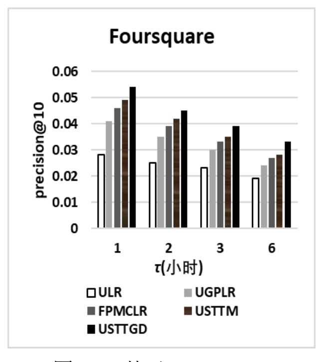


图 10　基于 Foursquare 的准确率实验对比图

Fig. 10 Precision experimental comparison based on Foursquare

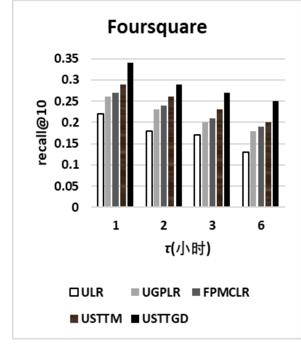


图 11　基于 Foursquare 的召回率实验对比图

Fig. 11 Recalling rate experimental comparison based on Foursquare

表 7　Gowalla 数据集中不同 τ 值下的推荐性能对比表

Table 7　Table of recommendation performance comparisons for different τ values in Gowalla dataset

算法	评价指标	τ(小时)			
		1	2	3	6
ULR	precision@10	0.015	0.013	0.012	0.01
	recall@10	0.23	0.17	0.12	0.07
UGPLR	precision@10	0.038	0.035	0.033	0.031
	recall@10	0.39	0.28	0.2	0.15
FPMCLR	precision@10	0.047	0.043	0.038	0.036
	recall@10	0.45	0.37	0.25	0.19
USTTM	precision@10	0.048	0.046	0.039	0.038
	recall@10	0.47	0.37	0.26	0.21
USTTGD	precision@10	0.054	0.05	0.043	0.041
	recall@10	0.52	0.41	0.3	0.26

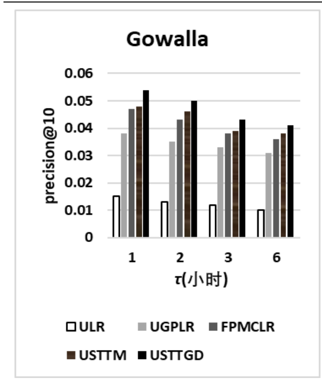


图 12　基于 Gowalla 的准确率实验对比图

Fig. 12 Precision experimental comparison based on Gowalla

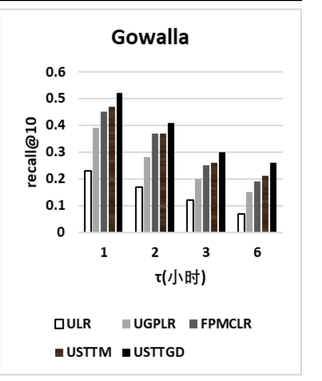


图 13　基于 Gowalla 的召回率实验对比图

Fig. 13 Recalling rate experimental comparison based on Gowalla

通过实验 2 结果可以看出，随着时间阈值(τ)约束条件的增加，各类算法的推荐精度和召回率均会降低。这是由于当时间间隔增大时，用户可能会移动到离当前所在的兴趣点较远的位置，从而降低连续兴趣点推荐的性能。因此，控制好时间间隔对于模型的预测效果也有着重要的意义。

5　结束语

本文利用用户在 LBSN 中的历史签到记录，提出了一种统一兴趣点推荐模型，首先根据时间分割 LDA 模型对用户语义位置信息建模，接着引入时间变量可在单个 LDA 模型中生成时间主题，缓解了之前的数据稀疏性问题。然后通过建立网格来结合区域影响因素和兴趣点间连续过渡因素，最终向用户产生推荐。实验结果表明，本文所提新型模型有效

地融合了多种情境因素，并且在各项性能评价指标上均优于现有的主流推荐算法。下一步的研究工作是深入挖掘用户签到数据中的其他属性，以达到更优的推荐效果。

参考文献：

[1] Chen Lingjiao, Gao Jian.A trust-based recommendation method using network diffusion processes [J]. Physica A: Statistical Mechanics and its Applications, 2018, 506(9): 679-691.

[2] Yang Carl, Bai Xiaolan, Zhang Chao, *et al.* Bridging collaborative filtering and semi-supervised learning: a neural approach for POI recommendation[C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Neew York:ACM Press,2017

[3] Zhang Dachuan, Li Mei, Wang Changdong.Point of interest recommendation with social and geographical influence[C]//Proc of IEEE International Conference on Big Data. 2017

[4] Hu Bo, Mohsen Jamali, Martin Ester.Spatio-temporal topic modeling in mobile social media for location recommendation[C]//Proc of the 13th International Conference on Data Mining. 2013:1073-1078.

[5] Li Huayu, Ge Yong, Hong Richang, *et al.* point-of-interest recommendations: learning potential check-ins from friends[C]//Proc of the 22 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. NewYork: ACM Press, 2016: 975-984.

[6] Li Huayu, Hong Richang, Wu Zhiang, *et al.* Spatial-temporal probabilistic matrix factorization for point-of-interest recommendation[C]//Proc of SIAM International Conference on Data Mining. 2016:117–125.

[7] Ference G, Ye Mao, Lee Wangchien. Location recommendation for out-of-town users in location-based social networks[C]//Proc of the 22nd ACM international conference on Information & Knowledge Management. New York:ACM Press, 2013:721–726.

[8] Yin Hongzhi, Zhou Xiaofang, Shao Yingxia, *et al.* Joint modeling of user behaviors for point-of-interest recommendation[C]//Proc of the 24th ACM International on Conference on Information and Management. New York:ACM Press, 2015:1631–1640.

[9] Wang Weiqing, Yin Hongzhi, Chen Ling, et al, Geo-sage: a geographical sparse additive generate model for spatial item recommendation[C]//Proc of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2015:125–126.

[10] Leskovec J, Krevl A. Snap datasets: stanford large network dataset collect[DB/OL]. (2014-06). http://snap. stanford. edu/data.

[11] He Jing, Li Xin, Liao Lejian, *et al.* Inferring a personalized next point-of-interest recommendation model with latent behavior patterns[C]//Proc of the 30th AAAI Conference on Artificial Intelligence. 2016.

[12] Yang Dingqi, Zhang Daqing, Zheng V W, et al, Modeling activity preference by leveraging spatial-temporal characteristic[J]. IEEE Trans on System Man and Cybernetics, Systems, 2015,45(1):129-142.

[13] 任星怡，宋美娜，宋俊德. 基于用户签到行为的兴趣点推荐 [J]. 计算机学报, 2017, 35(1): 28-51. (Ren Xingyi, Song Meina, Song Junde. Recommendation of points of interest based on user check-in behavior [J]. Journal of Computer Science, 2017, 35(1): 28-51.)

[14] 袁适之，李晶，李石君,等. 一种基于位置社交网络的地点推荐算法

[J]. 计算机应用研究, 2016, 33(7): 2003-2006. (Yuan Shizhi, Li Jing, Li Shijun, *et al*. A location recommendation algorithm based on location social network [J]. Application Research of Computers, 2016,33(7): 2003-2006.)

[15] 任看看, 钱雪忠. 协同过滤算法中的用户相似性度量方法研究 [J]. 计算机工程, 2015, 41(8): 18-22. (Ren Kankan, Qian Xuezhong. User similarity measurement in collaborative filtering algorithms [J]. Computer Engineering, 2015, 41(8): 18-22.)

[16] Xie Wenlei, Bindel D, Demers A, *et al*. Edge-weighted personalized PageRank: breaking a decade performance barrier[C]//Proc of ACM International Conference on Knowledge Discovery and Data Mining, New York:ACM Press, 2015.

[17] Zhang Wei,Wang Jianyong. Location and time aware social collaborative retrieval for successive point-of-interest recommendation[C]//Proc of the 24th ACM International on Conference on Information and Knowledge Management.New York:ACM Press, 2015:1221–1230.

[18] Zhao Shenglin, Zhao Tong, Yang Haiqin, *et al*. Stellar: spatial-temporal latent ranking for successive point-of-interest recommendation[C]//Proc of the 30th AAAI Conference on Artificial Intelligence, 2016.

[19] Lichman M, Smyth P.Modeling human location data with mixtures of kernel densities[C]//Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York: ACM Press, 2014,:35–44.